Optimal Sink Deployment for Distributed Sensing of Spatially Nonstationary Phenomena

Lorenzo A. Rossi, Bhaskar Krishnamachari and C.-C. Jay Kuo

Ming Hsieh Department of Electrical Engineering
University of Southern California, Los Angeles, CA 90089-2564, USA.
E-mail: lrossi@usc.edu, bkrishna@usc.edu, cckuo@sipi.usc.edu

Abstract—The optimal deployment of sinks in a sensor region for power efficient data gathering of a physical phenomenon is investigated in this work. In the system of consideration, nodes perform lossless distributed coding of the sensed data and the spatial statistics of the monitored phenomenon are possibly nonstationary due to heterogeneity in the sensing environment (e.g., variations in the altimetric profile). Non-stationary spatial statistics lead to uneven spatial profiles of the bit rates, unlike stationary statistics. The properties of rate profiles and the consequent optimal sink locations for a broad class of spatially non-stationary covariance models are studied by mathematical analysis and numerical examples.

Index Terms—Correlated data gathering, spatially non-stationary covariance models, deployment.

I. INTRODUCTION

The topic of energy efficient acquisition of a spatio-temporal phenomenon via wireless sensor networks, namely, correlated data gathering, is investigated in this work. In a data gathering round, each node aggregates its own sensed data with data from other nodes, performing entropy coding, and forwards them to its parent node. Final destinations for all data packets are nodes called sinks. The overall energy needed by the network for gathering data depends on the total length of the set of routes to sinks, which is called the data gathering tree, and on the rates generated by nodes. The rate depends on the spatial correlation among the data.

In the sensor network literature, problems of power efficient data gathering under an implicit (or explicit) assumption of spatially stationary statistics have been extensively studied, e.g., [11], [8], [3]. With few and very recent exceptions [7], spatial covariances are assumed to be stationary and isotropic. However, real phenomena are typically characterized by nonstationary spatial covariance models due to heterogeneity (e.g., variations in the altimetric profile) in the underlying environment [10]. So far, little is known on the specific impact of nonstationarity on correlated data gathering. Nonstationary models for geophysical phenomena have been studied with an increasing interest in the field of geostatistics [10], [12].

Focus of this work is the problem of optimal deployment of sinks for power efficient data gathering of a spatially nonstationary phenomenon. An evident consequence of spatial nonstationarity is that the rate profile has an uneven (or non-uniform) spatial distribution. Some areas may be characterized by more correlated data, therefore generating lower rates, while other areas may generate higher rates due to higher spatial harshness of the monitored phenomenon. Intuitively, if the rate profile has an uneven distribution with respect to the region centre, placing the sink closer to the higher rate area could be more energy efficient than keeping it at the region centre. Under the assumption of an even spatial distribution of rates generated by sensor nodes, the center would be the most energy efficient location, which is true when the monitored phenomenon is characterized by stationary spatial statistics. The optimal sink placement problem under spatial nonstationarity of the phenomenon is however not well studied in the literature. The problem of multisink deployment was addressed in [6], where each node was assumed to generate a constant rate under a regular grid topology.

We address this problem by studying the properties of rate profiles for a general class of nonstationary spatial covariance models both analytically and numerically. An interesting aspect is that we are able to relate geo-statistical properties of a physical phenomenon to rate profiles and, hence, to specific properties of the sensor network. The main challenges arise from the complexity of nonstationary spatial covariance models and the multiplicity of factors that affect the rate profile (e.g., the type of the distributed source coding scheme being adopted, the structure of the data gathering tree). Our main result is that the optimal location of the sink turns out to be close (usually within one hop) to the centre of the sensor region for a general class of nonstationary covariance models even though preliminary intuition would have suggested that the optimal location may be significantly distant from the centre.

The rest of this document is organized as follows. Some background on correlated data gathering is given in Section II. Main assumptions are made and the problem is formulated in Sec. III. Sec. IV presents analytical results on sink placement under the assumption of densely deployed sensor networks. Analytical results on the spatial profile of the rate are given in Sec. V. Experimental results are shown in Sec. VI and conclusions are given in Section VII.

II. BACKGROUND

Sensor networks deal with sampling of continuous quantities. In this work, nodes are assumed to perform uniform quantization with step size equal to $\Delta$. If a set of $n$ random variable $\{Z_i\}$ is characterized by a multivariate normal distribution $\mathcal{N}_n(\mu, K)$, where $\mu$ is the vector of the means and
$K$ is the covariance matrix, then the joint entropy of $\{Z_i\}$ is approximately [1]:

$$H(Z_1, Z_2, \ldots, Z_n) \approx \frac{1}{2} \log_2(2\pi e)^n |K| - n \log_2 \Delta. \quad (1)$$

The two main distributed lossless coding approaches studied in sensor networks are [3]: explicit communication and Slepian-Wolf. In this paper, we consider explicit communication although our results hold or can be easily extended to the Slepian-Wolf. In the explicit communication scheme, a node encodes its data conditioning on the data received from the children. The amount of side information available at a certain node $i$ depends on the number of children of $i$ and on the spatial correlation among the data of the children.

**Example 1:** In Figure 1, node $Z_2$ encodes its own data conditioning on the ones from node $Z_3$ and then send them to node $Z_1$. The incremental rate at $Z_2$ is at least $H(Z_2/Z_3)$; the incremental rate at $Z_1$ is at least $H(Z_1/Z_2, Z_3)$. The total rate at node $Z_1$ is at least: $R(1) = H(Z_3) + H(Z_2/Z_1) + H(Z_1/Z_2, Z_3)$.

![Fig. 1. An example of data gathering tree.](image)

**III. Problem Formulation**

We consider a set of $N$ sensor nodes deployed over a compact region $\Omega \subseteq \mathbb{R}^p$, sampling a spatio-temporal process $Z(x,t)$, $x \in \Omega$, $t \in \mathbb{R}$ characterized by a separable spatial covariance model $K(Z(x_1,t_1), Z(x_2,t_1))$. All the nodes are connected via a tree structure, the data gathering tree, to a special node called sink, the root of the tree. The samples of the field $Z(.)$ are transmitted to the sink through the data gathering tree. Our goal is to find the location $s$ of the sink in $\Omega$ that minimizes the overall network cost of transmitting the data to the sink.

We assume that a node performs lossless compression of its own sensing data via an explicit communication scheme, i.e., using the side information from the children. See [3]. We adopt the shortest path tree (SPT) as data gathering structure. The total amount of information transmitted by a node $i$, including the data received from its children, is the rate, $R(i)$. We define the incremental rate, $r(i)$, as the amount of information generated by node $i$ encoding its own sensing data. Let $T_i$ be the tree rooted at node $i$, then $r(i)$ is given by: $r(i) = R(i) - \sum_{j \in T_i \setminus \{i\}} R(j)$. Thus the incremental rate is the net amount of information at node $i$. For the explicit communication scheme, we have that $r(i) \geq H(Z_i/Z_j)$, $j \in T_i \setminus \{i\}$. We also assume that the sampling time at the sensor nodes is synchronized and that the network is faster than the phenomenon. Therefore effects of temporal correlation can be disregarded.

In general, the energy spent by each a node $i$, located at $x_i$, for transmitting a packet, depends on the rate $R(i)$ (i.e. the number of bits) transmitted and on the distance $d_i$ to the receiver of the packet. Therefore a typical expression for the total energy spent by network for a round of data gathering cost is [3]:

$$E = \sum_{i=1}^{N} R(i)d_i^k, \quad (2)$$

where $k = 2 - 4$.\(^\dagger\)

We derive an approximate expression for the energy cost (2) that makes explicit the dependency on the sink location and depends on the incremental rate.

**Proposition 1:** With the assumptions of fixed radio range communications and shortest path data gathering tree, an approximation up to a constant factor of the energy cost function is given by:

$$E' = \sum_{i=1}^{N} r(i)d(x_i, s), \quad (3)$$

$d(x_i, s)$ is the Euclidean distance between node and sink locations, i.e. $x_i$ and $s$.

**Proof:** Thanks to the fixed radio range assumption, we have:

$$E = \sum_{i=1}^{N} r(i) \sum_{j>i} n_{hop_i} l_{r_{ij}},$$

where $n_{hop_i}$ is the number of hops along the link from node $i$ to the sink. By approximating the number of hops with the Euclidean distance and disregarding multiplying constants, we prove that expression (3) is a valid cost function.

**IV. Massively Dense Sensor Networks**

The aim of this section is to find analytical solutions of problem (3) under some simplifying assumptions. We consider a massively dense sensor network performing correlated data gathering, so that the rate profile can be considered as a continuous function. We assume that the nodes are distributed over a compact region $\Omega \subseteq \mathbb{R}^p$, $p = 1, 2$, with spatial density $\delta(x)$. We denote the area (or, if $p = 1$, the length) of $\Omega$ as $|\Omega|$. We assume that the density $\delta(x)$ is such that the incremental rate can be assumed as a continuous function over $\Omega$. In this section, we denote the incremental rate as $r(x, s)$, to highlight its dependency on the sink location $s$.

Note that $r(x, s)$ depends also on the node density, $\delta(x)$, on the aggregation strategy chosen and on the covariance function $K(Z(x_1), Z(x_2)) = K(x_1, x_2)$. The incremental rate is nonnegative: $r(x, s) \geq 0$ everywhere in the sensor region, $\Omega$. It is measured in bits/m$^p$. The total amount of information generated over $\Omega$ can be expressed as:

$$I(\Omega, s) = \int_{\Omega} r(x, s) dx. \quad (4)$$

This is the total information collected at the sink after a data gathering round.

\(^\dagger\)In noisy wireless channels exp[R(i)] is considered instead of R(i) [9].
Based on Proposition 1, we express the cost of sending \( r(x) \) bits from \( x \) to the sink, at point \( s \), as \( dE = r(x)\|x-s\| \), where \( \|x-s\| \) is the Euclidean distance between the two points. Hence, the total cost of a data gathering round for the network can be written as:

\[
E(\Omega, s) = \int_{\Omega} r(x, s)\|x-s\|dx. \tag{5}
\]

Therefore the optimal sink location problem, under the assumption of massively dense sensor networks, is written as:

\[
s^* = \arg\min_s E(\Omega, s) \tag{6}
\]

In the multiple sink case, the data aggregation cost can be expressed as:

\[
E(\Omega, s_i) = \sum_i \int_{\Omega_i} r_i(x, s)\|x-s_i\|dx. \tag{7}
\]

Finding the minimal aggregation energy requires finding the optimal partition \( \{\Omega_i\} \) and then the optimal sink location within each subregion \( \Omega_i \), i.e.:

\[
\{\Omega_i^*, s_i^*\}_{i=1}^n = \arg\min_{\{\Omega_i, s_i\}} E(\Omega, s_i). \tag{8}
\]

We now show the convexity of the objective function (5).

**Proposition 2:** Consider a certain rate profile \( r(x, s) \) defined over a compact region \( \Omega \). The total information \( I(\Omega') \), \( \Omega' \subseteq \Omega \) is a monotonically increasing function of the monitored area.

**Proof:** If \( \Omega_1 \subseteq \Omega_2 \), then \( I(\Omega_1) \leq I(\Omega_2) \), because \( |\Omega_1| \leq |\Omega_2| \) and \( r(x, s) \geq 0 \).

A similar result holds for the energy:

**Lemma 1:** Consider a rate profile, \( r(x) \), defined over a compact region \( \Omega \). For a fixed sink location \( s \), the total aggregation energy \( E(\Omega, s) \) is a monotonically increasing function of the monitored area \( \Omega \).

**Proof:** It follows from the fact that both \( r(x, s) \) and \( \|s-x\| \) are nonnegative for every \( x \).

To further simplify the analysis, we disregard the dependency of \( r(.) \) on the sink location \( s \) for the rest of the section and denote it as \( r(x) \). This simplification is shown valid in Section V.

**Proposition 3:** If \( \{x_j : r(x_j) = 0\} \) has zero measure, then the objective function (5) is convex in \( \Omega \subset \mathbb{R} \).

**Proof:** Let \( \Omega = [0, L] \), by setting equal zero the first order derivative of (5) with respect to \( s \), we have:

\[
\int_0^{s_{opt}} r(x)dx = \int_x^L r(x)dx. \tag{9}
\]

This implies that the objective function has only one minimum point, because \( I(.) \) is a monotonically increasing function. Besides, it can be shown that the second order derivative of (5) is positive. Therefore the objective function is convex.

This means that in one dimension, the optimal location of the sink is such that the amount of information associated to the two subregions \([0, s^-] \) and \([s^+, L] \) is the same.

**Proposition 4:** The objective function (5) is convex in \( \Omega \subset \mathbb{R}^2 \).

**Proof:** The result can be proved by writing the Hessian of (5) and showing that it is positive definite for all \( s \in \Omega \subset \mathbb{R}^2 \).

**Example 2:** We consider here a slowly increasing rate profile, \( r(x) = \log_2(a+bx) \), with \( a, b > 0 \) and \( x \in [0, 1] \). For \( a = 2 \) and \( b = 10 \), the optimal sink location is \( s = .6 \). Besides, the ratio between the energy cost associated to the optimal sink location and the cost associated to the sink at the centre, \( E(.5)/E(.6) = .95 \). I.e., there is only a 5% energy saving in moving the sink from the centre to an optimal location. The value of the energy as function of the sink location is shown in Figure 2.

Fig. 2. Data aggregation cost as a function of the sink location.

V. ANALYSIS OF THE RATE PROFILE

In this section, we study the relationship between correlation model and spatial profile of the incremental rate, \( r(x, s) \), \( x, s \in \Omega \), analyzing in particular the impact of factors such as the spatial covariance of the monitored phenomenon, and the location of the sink. To this end, we need some further assumptions on the structure of the spatial correlation model.

We introduce a general class of nonstationary covariance models (Subsec. V-A) characterized by slow variability over space. Then, we analyze the sensitivity of \( r(.) \) to the sink location and study the steepness of the rate profile (Subsec. V-C).

We assume that the phenomenon \( Z \) has a multivariate Gaussian spatial distribution \( N_\mu, \Sigma \), where \( n \) is the number of observation points and \( \Sigma \) is the covariance matrix. Thus the incremental rate \( r(i) \) at node \( i \) can be computed via the following expression [1]:

\[
r(i) := \frac{1}{2} \log_2(2\pi e) \frac{\det(\Sigma[T_i])}{\det(\Sigma[T_i \setminus \{i\}])} - \log_2 \Delta, \tag{10}
\]

where \( \Delta \) is the quantization step size, \( \Sigma[T_i] \) is the covariance matrix of the subset of nodes \( T_i \). In the case of explicit communication, \( T_i \) is the subtree of nodes rooted at \( i \) and \( T_i \setminus \{i\} \) is the set of children of \( i \).

A. A class of nonstationary covariance models

Closed form expressions for nonstationary spatial covariances are complicated, because they obey the property of nonnegative definiteness [12]. We define a general class of covariances, characterized by a simple expression, by bounding ‘deviation from the stationarity’ of the models. Consider the two covariance coefficients \( K_1 = K(x, x+\Delta x) \) and \( K_2 =
We bound the absolute value of their difference, \(|K_1 - K_2|\), with a function growing linearly with the distance separating the two different pairs of measurements, i.e. \(|x - y|\), because we assume that a model for many nonstationary natural phenomena would still exhibit a quasi-stationary behavior on the short range [10]. If \(K()\) is isotropic and stationary \(K_1 = K_2\). We then introduce an additional condition to express that as the distance between two measurement points increases, their covariance decreases, for instance with exponential decay [12].

Therefore, for the points \(x, y, w \in \Omega \subseteq \mathbb{R}^p\) and the vector \(\Delta x \in \mathbb{R}^p\), we consider covariance functions obeying the following additional conditions on short and long range behavior:

\[
|K(x, x + \Delta x) - K(y, y + \Delta x)| \leq a|x - y|, \tag{11}
\]

\[
K(x, w) \leq be^{-c|x - w|}, \tag{12}
\]

for some real constants \(a, b\) and \(c\). We refer to condition (11) as distance dependent variability and to condition (12) as exponential decay. These conditions seem valid for many natural phenomena and are respected by some existing nonstationary correlation models [2], [5].

**B. Variability with respect to the sink location**

From numerical results, we have observed that the dependency of the rate profile from the sink location is usually fairly weak and that therefore, given two distinct sink locations, \(s_1\) and \(s_2\), we can consider \(r(x, s_1) \approx r(x, s_2)\) for the purpose of optimal sink placement in \(\Omega\).

To gain better insights on the behavior of the profile of the incremental rate, we want to find the following upper-bound

\[
\max_{x_1, x_2, x_i} |r(x_i, s_1) - r(x_i, s_2)| \quad s_1, s_2, x_i \in \Omega, \tag{13}
\]

under the conditions (11-12) on the correlation model and for the explicit communication case. To compute the upper bound of the expression (13), we assume the worst case scenario: i.e. \(s_1, s_2, x_i\) are such that the set of children of the node at \(x_i\) changes completely as the sink switches from \(s_1\) to \(s_2\). See Figure 3. We have that, due to condition (12), only a limited number \(n\) of children is used as the others have little correlation with point \(x_i\). Hence

\[
|r(x_i, s_1) - r(x_i, s_2)| = \frac{1}{2} \log_2(2\pi e) \left| \frac{K^{(1)}_n || K^{(2)}_{n-1} || K^{(1)}_{n-1} || K^{(2)}_n }{K^{(1)}_n || K^{(2)}_n} \right|, \tag{14}
\]

where \(K^{(j)}_{n-1}\) is the covariance matrix associated to children of node \(i\) as the sink is \(s_j\) and \(K^{(j)}_n\) is the covariance matrix associated to node \(i\) and its children. We express the matrices at the numerator as perturbations of the matrices at the denominator: \(K^{(1)}_n = K^{(2)}_n + M_n\). We have that \(\|M_n\|\) is bounded due to condition (11). By expressing the determinants as products of eigenvalues and applying the Bauer-Fike theorem, from matrix perturbation theory [4], we have that

\[
\max_{s_1, s_2, x_i} |r(x_i, s_1) - r(x_i, s_2)| \leq \log_2 \left( \frac{\lambda_{\max}^{(1)} + \delta_n^{(n)}(\lambda_{\max}^{(2)} + \delta_{n-1})^{n-1}}{\prod_{j=1}^{n} \lambda_j^{(2)} \prod_{j=1}^{n-1} \lambda_j^{(1)}} \right), \tag{15}
\]

where \(\delta_n = \|M_n\|\), \(\lambda_j^{(m)}\) is an eigenvalue of \(K^{(n)}\).

Small perturbations over a symmetric matrix lead to small variations of the eigenvalues and hence of the determinant. A covariance matrix is symmetric by definition. Hence the maximum variation of an eigenvalue is bounded by only a norm of the perturbation matrix, which is a small number because of the hypothesis on the slow variation of the covariance model (11). Similar results can be obtained for the case of perturbation of the node locations.

**C. Maximum steepness of the incremental rate**

If the incremental rate field has a nearly flat or an even profile, then the optimal location for the sink is the centre of the region. Hence, we want to evaluate how steep can be the profile \(r(.)\) and thus how eccentric can be the optimal sink location for a correlation structure obeying to the conditions of smooth spatial nonstationarity given in (11-12).

For simplicity, we consider a simplified scenario of regular one dimensional placement, where data are correlated only among close neighbors: i.e. \(E[x_i, x_j] = 0\) if \(|i - j| > 1\). We want to evaluate an upper bound for the expression

\[
|r(x_i) - r(x_{i+1})| = \left| \log_2(2\pi e) \left| \frac{K^{(1)}_n || K^{(2)}_{n-1} || K^{(1)}_{n-1} || K^{(2)}_n }{K^{(1)}_n || K^{(2)}_n} \right| \right|, \tag{16}
\]

By applying condition (11) and the Bauer-Fike theorem, an upper-bound for (16) is found to be

\[
|r(x_i) - r(x_{i+1})| \leq \log_2(2\pi e) \left| \frac{\sigma_x^2 + a \Delta x (\lambda_1 + \delta_2)^2}{\sigma_x^2 \lambda_1 \lambda_2} \right|^2, \tag{17}
\]

where \(\lambda_1 \geq \lambda_2\) are the eigenvalues of \(K^{(n)}\). Therefore, the rate profile is bounded by a logarithmic profile. We showed in Example 2 that this implies that the optimal sink location is generally close to the centre and, therefore, the amount of energy saved by optimally placing a sink is relatively low with respect to placing the sink at the centre.

**VI. EXPERIMENTAL RESULTS**

The rain fall data set provided by the University of Washington [13] was examined in our experiment. The data set provides the daily rain concentration in the North Western region of the US (Washington) for a period of 49 years. The observation points are located on regular grid, with minimum distance
between two points equal to $50 \text{Km}$. To remove the seasonality and estimate the spatial covariance, we considered the 365 day difference of the data (i.e. $W(x, t) := Z(x, t) - Z(x, t-365)$) and then the 10 day cumulative over a period of 80 days during the rainy season for a period of 30 years. Thus the spatial covariance between measurements at the locations $x_i$ and $x_j$ is estimated as $\hat{K}(x_i, x_j) = \frac{1}{T} \sum_{t=1}^{T} Z(x_i, t)Z(x_j, t)$. The resulting covariance estimates showed spatial non-stationarity especially in subregions characterized by mixed terrain (e.g. plains and mountains). We computed the rate profiles from the covariance estimates. The data were assumed to have a multivariate Gaussian distribution. Overall the rate profiles showed a fairly even distribution. We determined the optimal sink location computing the data aggregation cost (3) for all the grid locations within a subregion. For different subsets of the rain data, the optimal sink locations turn out to be within one hop from the centre of the regions being considered.

Example 3: We considered the subregion of the observation field shown in the box in Fig. 4. The rate profiles are shown in Fig. 5. The optimal sink location resulted to be less than one hop to the centre of the region (Fig. 4).

VII. CONCLUSION

We studied the problem of the energy efficient sink deployment for correlated data gathering. The fundamental assumption here is that the physical field being monitored is characterized by a nonstationary spatial covariance model. Nonstationarity leads to an uneven spatial distribution of the rates assigned to the sensor nodes and potentially to an optimal sink location different from the centre of the sensor region, unlike typically in the case of stationary spatial statistics. We provided analytical results on the optimal sink location and studied the properties of the spatial profiles of the rates for a general class of nonstationary spatial covariance models. Our main conclusion, validated on a real data set as well, is that the optimal location of the sink turns out to be close (usually within one hop) to the centre of the sensor region. This result goes against preliminary intuition on the impact of nonstationarity on correlated data gathering.

We plan to provide more analytical results and study the problem under different energy metrics. The broad goal of this research is to get a deeper understanding of the impact of more accurate modeling of spatio temporal phenomena on wireless sensor network problems and algorithms.

REFERENCES